Multiple breath washout quality control in the clinical setting

Frauchiger, Bettina S ; Carlens, Julia ; Herger, Andreas ; Moeller, Alexander ; Latzin, Philipp ; Ramsey, Kathryn A

Abstract: BACKGROUND Multiple breath washout (MBW) is increasingly used in the clinical assessment of patients with cystic fibrosis (CF). Guidelines for MBW quality control (QC) were developed primarily for retrospective assessment and central overreading. We assessed whether real-time QC of MBW data during the measurement improves test acceptability in the clinical setting. METHODS We implemented standardized real-time QC and reporting of MBW data at the time of the measurement in the clinical pediatric lung function laboratory in Bern, Switzerland in children with CF aged 4-18 years. We assessed MBW test acceptability before (31 tests; 89 trials) and after (32 tests; 96 trials) implementation of real-time QC and compared agreement between reviewers. Further, we assessed the implementation of real-time QC at a secondary center in Zurich, Switzerland. RESULTS Before implementation of real-time QC in Bern, only 58% of clinical MBW tests were deemed acceptable following retrospective QC by an experienced reviewer. After implementation of real-time QC, MBW test acceptability improved to 75% in Bern. In Zurich, after implementation of real-time QC, test acceptability improved from 38% to 70%. Further, the agreement between MBW operators and an experienced reviewer for test acceptability was 84% in Bern and 93% in Zurich. CONCLUSION Real-time QC of MBW data at the time of measurement is feasible in the clinical setting and results in improved test acceptability.

DOI: https://doi.org/10.1002/ppul.25119

Multiple breath washout quality control in the clinical setting

Bettina S. Frauchiger[1*], MD; Julia Carlens[1,2*], MD; Andreas Herger[3], BSc; Alexander Moeller[3], MD; Philipp Latzin[1], Ph.D; Kathryn A. Ramsey[1#], Ph.D.

1. Pediatric Respiratory Medicine, Department of Pediatrics, Inselspital, Bern University Hospital, University of Bern, Freiburgstrasse 8, 3010 Bern, Switzerland†

2. Clinic for Paediatric Pneumology, Allergology and Neonatology, Hannover Medical School, Carl-Neuberg-Strasse 1, 30625 Hannover, Germany

3. Division of Respiratory Medicine and Children's Research Center, University Children's Hospital Zurich, Steinwiesstrasse 75, 8032 Zurich, Switzerland †

† Institutions where research was mainly conducted

* Co-first authors: Authors contributed equally

[#] Corresponding author:
Kathryn Ramsey
Inselspital, Bern University Hospital,
Freiburgstrasse 8, Bern 3010 Bern, Switzerland
Email: Kathryn.ramsey@extern.insel.ch

Funding: This project was funded by the Swiss National Science Foundation, Grant Nr. PZ00P3_168173 (Latzin) / 1; 32003B_182719 (Ramsey)

Disclosures: Dr. Latzin reports grants from the Swiss National Science Foundation, during the conduct of the study; personal fees from Vertex, personal fees from Novartis, personal fees from Roche, personal fees from Polyphor, personal fees from Vifor, personal fees from Gilead, personal fees from Schwabe, personal fees from Zambon, personal fees from Santhera, grants from Vertex, outside the submitted work. Dr. Ramsey reports grants from the Swiss National Science Foundation during the conduct of the study and grants from Vertex, outside the submitted work.

Keywords: Multiple breath washout, lung clearance index, quality control, cystic fibrosis, children

Running head: Quality control for multiple breath washout

Wordcount (without abstract): 3355 words

Abstract: 220 words

Author contributions: BF, JC, PL, AM, and KR were responsible for the conception and design of this study. BF, JC, and KR drafted the quality control guidelines and the structured implementation in clinical routine. Data acquisition was conducted by BF and AH. BF, JC, AM, PL, and KR were responsible for data interpretation. Statistical analysis was conducted by BF

and KR. BF, JC, AM, PL, and KR drafted the manuscript and all authors revised and approved the manuscript for intellectual content before submission.

Abstract

Background: Multiple breath washout (MBW) is increasingly used in the clinical assessment of patients with cystic fibrosis (CF). Guidelines for MBW quality control (QC) were developed primarily for retrospective assessment and central overreading. We assessed whether real-time QC of MBW data during the measurement improves test acceptability in the clinical setting.

Methods: We implemented standardized real-time QC and reporting of MBW data at the time of the measurement in the clinical pediatric lung function laboratory in Bern, Switzerland in children with CF aged 4-18 years. We assessed MBW test acceptability before (31 tests; 89 trials) and after (32 tests; 96 trials) implementation of real-time QC and compared agreement between reviewers. Further, we assessed the implementation of real-time QC at a secondary center in Zurich, Switzerland.

Results: Before implementation of real-time QC in Bern, only 58% of clinical MBW tests were deemed acceptable following retrospective QC by an experienced reviewer. After implementation of real-time QC, MBW test acceptability improved to 75% in Bern. In Zurich, after implementation of real-time QC, test acceptability improved from 38% to 70%. Further, the agreement between MBW operators and an experienced reviewer for test acceptability was 84% in Bern and 93% in Zurich.

Conclusion: Real-time QC of MBW data at the time of measurement is feasible in the clinical setting and results in improved test acceptability.

1. Introduction

The lung clearance index (LCI) derived from the multiple breath washout technique (MBW) is sensitive to detect early lung disease in patients with cystic fibrosis (CF) [1-4]. With the availability of commercial MBW devices, LCI is increasingly being used as an outcome in routine clinical surveillance [5-10]. While MBW testing requires minimal cooperation from the subject, an acceptable test requires relaxed tidal breathing and a leak-free system[11], which can be challenging in young children and individuals with respiratory disease[12]. Besides, prospective quality control (QC) of MBW measurements can be challenging in the busy clinical setting.

Quality control guidelines for MBW focus primarily on retrospective analysis and central overreading of MBW measurements by experienced users for research studies and clinical trials [11,13-15]. The 2013 European Respiratory Society (ERS) and American Thoracic Society (ATS) consensus statement for inert gas washout measurements proposed initial recommendations for testing procedure and technical acceptability criteria [11]. Further to this, ATS published additional guidelines for the preschool age group [14]. Jensen *et al.* proposed comprehensive guidelines for retrospective quality control of MBW measurements, which involved both qualitative and quantitative criteria for trial grading and acceptability[13]. These guidelines were further implemented in a standardized MBW training and quality control platform for central overreading in clinical trials [15]. However, for LCI to be used as a clinical outcome, prospective reporting of acceptability and test results is required for clinical decision making.

Therefore, we aimed to implement prospective, real-time quality control of MBW measurements in the clinical pediatric lung function laboratory in Bern, Switzerland. The first aim was to evaluate the acceptability of MBW measurements performed in routine outpatient clinics in children with CF before and after the implementation of real-time quality control.

The second aim was to assess the implementation of real-time quality control of MBW measurements in a pediatric lung function laboratory with less experience in MBW testing (Zurich, Switzerland). The third aim was to evaluate agreement in MBW test acceptability between the operator and a retrospective reviewer.

2. Methods

2.1 Development of MBW quality control criteria

The quality control criteria used in this study were based on the ATS/ERS consensus statement guidelines, ATS pre-school MBW technical statement, and the publications by Jensen *et al* and Saunders *et al.* [11,13-15]. We used these guidelines to create a simplified matrix for qualitative assessment of Nitrogen (N_2) MBW measurements performed in routine clinical testing that can be applied at the time of the measurement and did not require any further retrospective assessment.

Our quality control criteria are presented in Table 1 and details of how our criteria differ from the ERS/ATS consensus statement are provided in Supplemental Table E1. Detailed instructions on how to apply the guidelines are presented in the online supplemental. An A grade represents a perfect trial with relaxed, regular tidal breathing throughout the measurement, a B grade represents a good quality trial with only minimal deviations, and a C grade represents an acceptable trial with moderate deviations but no highly abnormal breaths during the pre-phase or start of washout. A, B, and C grade trials are considered acceptable for outcome reporting. D grade represents trials with questionable quality due to variable breathing patterns, abnormal breaths, or evidence for hypo- or hyper-ventilation. D grade trials have no signs of leak and satisfy both the start and end of test criteria. However, D grade trials contain highly variable breathing patterns or abnormal breaths that might influence MBW results. While a clear consensus on how to handle trials with questionable breathing patterns has not been reached, we recommend rejection of D grade trials in our criteria. An F grade represents trials that need to be rejected due to not meeting the technical acceptability criteria for MBW: 1) Start of test criteria not met (last three breaths of pre-phase with normalized end-tidal N_2 concentration ≥ 77%); 2) End of test criteria not met (three

consecutive tidal breaths with normalized end-tidal N_2 concentration < 2.5%; 3) No evidence of leaks (for detailed instruction see online supplemental).

The overall test occasion is classified as acceptable when at least two trials are graded as acceptable (A, B, or C). We used the overall test repeatability criteria described in the consensus document (i.e. FRC variability within 25%)[11]. MBW outcomes from acceptable and repeatable test occasions are reported as the mean from all acceptable trials.

2.2 MBW data collection and study population

The N_2MBW measurements were collected using the Exhalyzer D device (Eco Medics, Duernten, Switzerland) with Spiroware software (version 3.2.1) and were performed according to international guidelines in both centers [11]. We approached all pediatric patients with CF attending their regular three monthly outpatient clinic visits aged 4 to 18 years. Approval was obtained from the local ethics committee in Bern. Patients and caregivers gave informed consent.

2.3 Test acceptability before implementation of real-time MBW quality control in Bern

Before implementation of real-time MBW quality control criteria into clinical routine in our centre, MBW operators were trained in data collection and general test acceptability. However, due to time restrictions, they were not required to perform a detailed assessment of test quality during the measurement. Operators occasionally excluded trials for outcome reporting in the software but did not routinely mark trial classification or test acceptability on the lung function reports. To assess the quality of these MBW measurements, 31 clinic visits from children with CF aged five to 18 years were evaluated. The visits were randomly selected by an independent person not involved in this study and only one visit per patient was included in the analysis. Retrospective quality control was performed by an experienced

reviewer who was involved in the development of the criteria and was blinded to any test comments by the MBW operator. The reviewer graded each trial individually and then assessed overall test acceptability.

2.4 Test acceptability after implementation of real-time MBW quality control in Bern

To implement real-time quality control of MBW measurements, operators in our center received instruction on how to perform quality control. A printed copy of the quality control criteria matrix was provided and operators were given a presentation on how to use the matrix, grade individual trials, determine test acceptability and repeatability, and report outcomes (detailed information provided in online supplement). All MBW operators were required to perform real-time quality control on all MBW measurements. The operators reported a grade for each trial and provided a standardized comment regarding the acceptability of the test occasion in the clinical report (example provided in the online supplement).

To assess the quality of MBW measurements after implementation of real-time quality control in routine clinical testing, 32 clinic visits from the same population of children with CF were evaluated. The visits were randomly selected by an independent person not involved in this study and only one visit per child was assessed. The experienced reviewer performed retrospective quality control of these measurements while being blinded to the real-time quality control assessment of the operator.

2.5 Assessment of real-time MBW quality control in a secondary centre in Zurich

To validate our findings, we implemented real-time MBW quality control in a centre (Zurich, Switzerland) with less experience in MBW methodology. We first assessed MBW measurements from a random subset of 34 clinical visits from pediatric patients with CF

attending their routine care in Zurich, Switzerland. The MBW operators were then instructed on how to use the quality control matrix and given access to the supplementary teaching material (see online supplemental). Mandatory real-time MBW quality control and reporting of test acceptability were implemented for all measurements. We then retrospectively assessed test acceptability and agreement between the operator and reviewer in a random subset of 30 MBW measurements.

2.6 Outcomes before and after real-time quality control

To determine whether performing real-time quality control influenced outcomes, we compared mean outcomes (LCI and FRC) and variability (coefficient of variation; (CV)) for each test occasion before and after quality control. Before real-time quality control was implemented, we examined the differences in the outcomes printed on the clinical report to outcomes from trials deemed acceptable by the reviewer. After implementing real-time quality control, we compared differences in the outcomes reported prospectively by the operator vs retrospectively assessed by the reviewer.

We also performed a sensitivity analysis to examine the impact of accepting or rejecting D grade trials on outcome calculation and overall test acceptability.

2.7 Data analysis

For this study, we examined MBW test acceptability before and after implementation of real-time quality control in the clinical setting. Before the implementation of real-time quality control, we compared test acceptability reported in the clinical reports generated by MBW operators to those reported following retrospective quality control by the experienced reviewer (BF). While trial classification was not documented on the printed report before implementation of our guidelines, the trials selected for outcome reporting were visible in the

spiroware software. After the implementation of real-time quality control, we compared test acceptability and trial acceptability reported at the time of the measurement by MBW operators to those reported following retrospective quality control by the experienced reviewer (BF). As a secondary outcome, we assessed the agreement in test acceptability and trial grading between the MBW operator and an experienced reviewer. Agreement was assessed using kappa statistics. We also compared MBW outcomes from accepted test occasions (mean and within-test coefficient of variation for FRC and LCI) reported by the operator and reviewer using unpaired t-tests. All statistical analysis was performed using Stata 16.0 (StataCorp 2019)[16].

3. Results

3.1 Study population

The demographic characteristics of study participants from both centers are summarized in Table 2. The population in Zurich was on average younger, however, the age range of patients was similar between both centers. Anthropometric characteristics (height, weight, BMI, and age) and MBW outcomes (LCI and FRC) were well matched between the two study populations.

3.2 MBW test and trial acceptability before implementing real-time quality control

MBW test acceptability results are summarized in Table 3. In Bern, before implementing real-time quality control, 89 MBW trials from 31 test occasions were evaluated. After retrospective analysis of quality control by the reviewer, 58% of test occasions were deemed acceptable. In terms of MBW trials, 51 (58%) were accepted, and 38 (42%) were rejected following retrospective quality control by the reviewer. The reasons for trial rejection (details provided in Table 4) included F grade trials whereby the technical acceptability criteria were not met and D grade trials with irregular breathing patterns.

In Zurich, before implementing real-time quality control, 97 MBW trials from 34 test occasions were evaluated. After retrospective quality control by the reviewer, only 38% of the test occasions were deemed acceptable. In terms of MBW trials, 44 (45%) were accepted and 53 (55%) were rejected following retrospective quality control by the reviewer. The reasons for trial exclusion (details provided in Table 4) were similar to Bern.

3.3 MBW test and trial acceptability after implementing real-time quality control

Test acceptability after implementing real-time MBW quality control is summarized in Table 3. In Bern, 96 trials from 32 MBW test occasions were evaluated. Test acceptability improved

from 58% (18/31) to 75% (24/32) and trial acceptability improved from 58% (51/89) to 69% (66/96) (Table 5). In Zurich, 91 trials from 30 MBW test occasions were evaluated after implementing real-time quality control. Test acceptability improved from 38% (13/34) to 70% (21/30), trial acceptability improved from 45% (44/97) to 67% (61/91) (Table 5).

3.4 Agreement between MBW operator and reviewer after implementing real-time quality control

After implementing real-time quality control, the agreement between operator and reviewer was high in both centers. For test acceptability agreement was 84% ($\kappa = 0.5$, $p < 0.001$) in Bern and 93% ($\kappa = 0.8$, $p = 0.001$) in Zurich. In Bern, four test occasions were rejected by the reviewer but not by the operator. The remaining four test occasions rejected by the reviewer were also rejected by the operator. In Zurich, only one test occasion was rejected by the reviewer but not by the operator, the remaining eight test occasions rejected by the reviewer were also rejected by the operator.

For trial grading, agreement was 68% ($\kappa = 0.6$, $p < 0.001$) in Bern and 73% ($\kappa = 0.6$, $p < 0.001$) in Zurich. When pooling the analysis to the comparison of acceptable (A-C grade), questionable (D grade) and rejected (F grade) trials, agreement in Bern was 88% ($\kappa = 0.7$, $p < 0.001$) and in Zurich 86% ($\kappa = 0.7$, $p < 0.001$) (Figure 1a and 1b). In Bern, operators identified all F grade trials, except for one which was given a D grade. In Zurich, 15/19 (80%) of the F grade trials were correctly identified by the operator, whereas two of the F grade trials were classified as D grade and two were classified as A-C grade by the operator. F grade trials consisted mostly of leaks, and the only trials in which the end of test criteria was not met were trials that were prematurely terminated by the operator. All the trials evaluated satisfied the start of test criteria, which meant that operators were consistently waiting enough time between trials.

3.5 MBW outcomes before and after quality control

We determined whether performing quality control significantly influenced MBW results. We found no differences in LCI or FRC mean values or variability reported by the MBW operator at the time of the measurement compared with the reviewer; for both periods, before the implementation of real-time quality control and after implementation (Supplemental Table E2 and E3). However, for individuals, performing quality control can have a minor impact on the outcomes (maximum change is +/- 0.5 LCI units), as shown in the Bland-Altman-Plot in the online supplemental Figure E1 and E2.

We performed a sensitivity analysis to determine the impact of accepting or rejecting questionable D grade trials. Accepting D grade trials increased overall test acceptability from 75% to 84% in Bern and from 70% to 90% in Zurich, and did not lead to significant differences in LCI and FRC (Supplemental Table E4).

4. Discussion

We assessed the quality of MBW measurements collected in clinical routine before and after implementing mandatory real-time quality control in two centers. We provided the MBW operators with a simplified quality control matrix according to current guidelines and assessed whether real-time quality control of MBW data during the measurement improves test acceptability in routine clinical testing. Following implementation of real-time quality control, acceptability of MBW measurements improved from 58% to 75% in Bern and from 38% to 70% in the validation center Zurich and resulted in excellent agreement between the operator and reviewer.

Implementing mandatory real-time quality control improved overall test acceptability and the ability of MBW operators to recognize and perform good quality MBW measurements in routine clinical testing. Before implementation of real-time quality control criteria, operators reported outcomes to the clinicians with trials that should have been rejected according to current quality control guidelines [11,13-15]. After implementing mandatory real-time quality control and providing the simplified guidelines in Bern, operators were able to correctly identify acceptable, questionable, and technically not acceptable trials with excellent agreement to an experienced reviewer. These findings were validated in a center with less experience in MBW measurements (Zurich) with similar agreement. We could show that performing real-time quality control in the clinical setting is feasible and that the improved compliance of our operators with the quality control guidelines resulted in higher test acceptability rates.

Our quality control guidelines are a simplified version of the current MBW consensus guidelines, preschool technical standards, and quality control guidelines by Jensen *et al.* and *Saunders et al*[11 14 13,15]. Due to the complexity and time-consuming nature of the MBW test,

there has been a focus on the need for detailed retrospective quality control and central over-reading by highly experienced MBW researchers. While central over-reading is a suitable approach for large, multi-centre research studies[15,17], MBW testing in routine outpatient clinics requires immediate reporting of outcomes. Therefore, real-time quality control by the operator is the only way to ensure that good quality MBW outcomes are used for clinical interpretation. Our quality control criteria (Table 1) provide simplified guidelines for MBW trial grading, trial acceptability, and test acceptability. They can be applied at the time of the measurement to allow immediate reporting of MBW data in clinics. We did not specifically assess if extra time was needed for reporting. However, MBW operators were encouraged to perform quality control during the waiting time between trials. The time-slots for MBW testing remained the same after implementing mandatory quality control, which indicates that not much extra time was necessary. To maintain high-quality MBW data in the clinical setting, we suggest that centres implement regular training sessions with operators, provide updates of recent literature in the field, and perform regular over-reading of random subsets of MBW measurements to provide feedback to the operators.

We found that systematic quality control of MBW measurements by an experienced reviewer did not lead to differences in mean LCI or FRC values compared with values reported without quality control. These findings are similar to those reported by Jensen *et al.*, who found no differences in mean LCI values following qualitative and quantitative review[13]. However, while both studies reported no significant differences in outcomes on a population level [18], reporting outcomes from technically not acceptable MBW trials can impact the outcomes of a test occasion on an individual level as we could show by comparing individual outcomes before and after quality control in supplemental Figures E1 and E2. In the clinical setting, longitudinal changes in MBW outcomes from one visit to the next are likely to influence treatment

decisions. Therefore, it is essential to ensure that only data from good quality MBW trials are reported.

In the clinical setting, trials that satisfy the technical acceptability criteria but have highly irregular breathing patterns (D grade trials) are a frequent challenge. There are limited data on the impact of non-tidal breaths and different degrees of tidal breathing variability on outcomes. However, the general consensus is that these questionable D grade trials should be rejected to avoid the potential impact on results and test variability [11,13-15]. We performed a sensitivity analysis and found that accepting D grade trials increased overall test acceptability without significantly influencing results. However, we found in general little differences in outcomes before and after the implementation of quality control. Further research will help to determine the impact of breathing pattern irregularities on MBW outcomes to inform quality control inclusion and exclusion criteria. However, to ensure reliability of MBW results for clinical reporting, we support the current recommendation to not accept questionable trials with irregular breathing patterns and perform strict quality control.

There are some limitations to our quality control criteria. Trial grading is based on qualitative criteria for breathing pattern that are inherently subjective. Users need to have familiarity with the MBW test to differentiate between minimally and moderately variable breathing pattern. However, many of these subjective criteria will only influence whether a trial is graded as either A, B or C, all of which are technically acceptable trials. Our criteria could have been simplified further to only include acceptable, questionable, or rejected trial grading, however, MBW operators stated that it was beneficial to be able to recognize what constitutes a technically perfect trial and understand which deviations can still be accepted. Further, our quality control criteria are only applicable to N_2MBW measurements. This study was

performed using Eco Medics equipment and Spiroware 3.2.1 software whereby flow-volume loops, nitrogen, oxygen, and carbon dioxide signals are visible during and after the measurement. It is unclear how easily these criteria can be applied to data collected in other devices, software, and alternative tracer gases.

Conclusion

Real-time quality control of MBW data at the time of the measurement is feasible in the clinical setting and results in improved test acceptability with excellent agreement between MBW operators and experienced reviewers. Our quality control criteria provide simplified guidelines for MBW trial grading, trial acceptability, test acceptability, and outcome reporting that can be applied in centers with limited experience in MBW methodology. Performing quality control at the time of the measurement ensures that only outcomes from good quality MBW trials are reported to the clinician.

Acknowledgments

The authors also thank to all our patients and families for allowing their MBW data to be used for research. The authors thank to all our lab technicians for collecting daily MBW measurements, especially Mrs. Lüscher, Mrs. Vessaz, Mrs. Krattinger and Mrs. Wirz in Bern as well as Mrs. Böhringer, Mrs. König, Mrs. Noser and Mr. Vogt in Zurich.

Conflicts of Interest

Dr. Latzin: personal fees from Vertex, Novartis, Roche, Polyphor, Vifor, Gilead, Schwabe, Zambon, Santhera, grants from Vertex, all outside this work. All other authors have no conflicts of interest.

References

1. Gustafsson PM, Aurora P, Lindblad A. 2003. Evaluation of ventilation maldistribution as an early indicator of lung disease in children with cystic fibrosis. Eur Respir J. 2003; 22(6):972.

2. Gustafsson PM, De Jong PA, Tiddens HA, Lindblad A. 2008. Multiple-breath inert gas washout and spirometry versus structural lung disease in cystic fibrosis. Thorax. 2008; 63(2):129-134.

3. Jensen R, Stanojevic S, Gibney K, Salazar JG, Gustafsson P, Subbarao P, Ratjen F. 2013. Multiple breath nitrogen washout: A feasible alternative to mass spectrometry. PLoS One. 2013; 8(2):e56868.

4. Owens CM, Aurora P, Stanojevic S, Bush A, Wade A, Oliver C, Calder A, Price J, Carr SB, Shankar A et al. 2011. Lung clearance index and hrct are complementary markers of lung abnormalities in young children with cf. Thorax. 2011; 66(6):481-488.

5. Aurora P, Stanojevic S, Wade A, Oliver C, Kozlowska W, Lum S, Bush A, Price J, Carr SB, Shankar A et al. 2011. Lung clearance index at 4 years predicts subsequent lung function in children with cystic fibrosis. Am J Respir Crit Care Med. 2011; 183(6):752-758.

6. Davies J, Sheridan H, Bell N, Cunningham S, Davis SD, Elborn JS, Milla CE, Starner TD, Weiner DJ, Lee PS et al. 2013. Assessment of clinical response to ivacaftor with lung clearance index in cystic fibrosis patients with a g551d-cftr mutation and preserved spirometry: A randomised controlled trial. Lancet Respir Med. 2013; 1(8):630-638.

7. Fuchs SI, Gappa M. 2011. Lung clearance index: Clinical and research applications in children. Paediatr Respir Rev. 2011; 12(4):264-270.

8. Kent L, Reix P, Innes JA, Zielen S, Le Bourgeois M, Braggion C, Lever S, Arets HG, Brownlee K, Bradley JM et al. 2014. Lung clearance index: Evidence for use in clinical trials in cystic fibrosis. J Cyst Fibros. 2014; 13(2):123-138.

9. Stanojevic S, Davis SD, Retsch-Bogart G, Webster H, Davis M, Johnson RC, Jensen R, Pizarro ME, Kane M, Clem CC et al. 2017. Progression of lung disease in preschool patients with cystic fibrosis. Am J Respir Crit Care Med. 2017; 195(9):1216-1225.

10. Fuchs SI, Eder J, Ellemunter H, Gappa M. 2009. Lung clearance index: Normal values, repeatability, and reproducibility in healthy children and adolescents. Pediatr Pulmonol. 2009; 44(12):1180-1185.

11. Robinson PD, Latzin P, Verbanck S, Hall GL, Horsley A, Gappa M, Thamrin C, Arets HGM, Aurora P, Fuchs SI et al. 2013. Consensus statement for inert gas washout measurement using multiple- and single- breath tests. Eur Respir J. 2013; 41(3):507.

12. Yammine S, Summermatter S, Singer F, Lauener R, Latzin P. 2016. Feasibility of nitrogen multiple-breath washout in inexperienced children younger than 7 years. Pediatr Pulmonol. 2016; 51(11):1183-1190.

13. Jensen R, Stanojevic S, Klingel M, Pizarro ME, Hall GL, Ramsey K, Foong R, Saunders C, Robinson PD, Webster H et al. 2016. A systematic approach to multiple breath nitrogen washout test quality. PLoS One. 2016; 11(6):e0157523.

14. Robinson PD, Latzin P, Ramsey KA, Stanojevic S, Aurora P, Davis SD, Gappa M, Hall GL, Horsley A, Jensen R et al. 2018. Preschool multiple-breath washout testing. An official american thoracic society technical statement. Am J Respir Crit Care Med. 2018; 197(5):e1-e19.

15. Saunders C, Jensen R, Robinson PD, Stanojevic S, Klingel M, Short C, Davies JC, Ratjen F. 2019. Integrating the multiple breath washout test into international multicentre trials. Journal of Cystic Fibrosis. 2019.
16. StataCorp. 2019. Stata statistical software: Release 16. In: LLC S, editor.: College Station, TX.
17. O'Neill K, Lakshmipathy GR, Ferguson K, Cosgrove D, Hill AT, Loebinger MR, Carroll M, Chalmers JD, Gatheral T, Johnson C et al. 2018. Quality control for multiple breath washout tests in multicentre bronchiectasis studies: Experiences from the bronch-uk clinimetrics study. Respir Med. 2018; 145:206-211.
18. Lenherr N, Ramsey KA, Jost K, Hornwall L, Singer F, Yammine S, Latzin P. 2018. Leaks during multiple-breath washout: Characterisation and influence on outcomes. ERJ Open Res. 2018; 4(1).

Tables

Table 1: Quality control criteria

A (perfect)	**Pre-phase**	Regular tidal breathing: relaxed, regular, and appropriate for size
	Start of washout	Regular tidal breathing: relaxed, regular, and appropriate for size
	Rest of washout	Regular tidal breathing: relaxed, regular, and appropriate for size
	EELV	Stable: no drift or step change
B (good)	**Pre-phase**	Regular tidal breathing: relaxed, regular, and mostly appropriate for size
	Start of washout	Regular tidal breathing: relaxed, regular, and mostly appropriate for size
	Rest of washout	Minimally variable breathing: flow minimally variable, irregular (small swallow/hesitation), V_T slightly outside target range
	EELV	Minimal to moderate drift; no step change
C (acceptable)	**Pre-phase**	Minimally variable breathing: flow minimally variable, irregular (small swallow/hesitation), V_T outside target range
	Start of washout	Minimally variable breathing: flow minimally variable, irregular (small swallow/hesitation), V_T outside target range
	Rest of washout	Moderately variable breathing: flow moderately variable, abnormal breath (sigh, cough, breath-hold) with no release of trapped gas, V_T outside target range
	EELV	Minimal to moderate drift; step change in rest of washout with no shift in N_2 signal
D (questionable)	**Pre-phase**	Moderately variable breathing: flow moderately variable, abnormal breath (sigh, cough, breath-hold), V_T outside target range
	Start of washout	Moderately variable breathing: flow moderately variable, abnormal breath (sigh, cough, breath-hold), V_T outside target range
	Rest of washout	Highly variable breathing: flow highly variable, forced expiration, abnormal breath (sigh, cough, breath-hold) with release of trapped gas, evidence for hypo/hyperventilation (CO_2 outside 4-6% range)
	EELV	Step change in rest of washout resulting in shift in N_2 signal
F (reject)	**Start of test criteria** (last three breaths of pre-phase N_2 concentration ≥ 77%) **not met**	

	End of test criteria (three consecutive tidal breaths with normalized N_2 concentration < 2.5%) **not met**	
	Leak during washout	
Overall test acceptability	**Minimum two acceptable trials**	
	A, B, C trials	Accept (if FRC variability within 25%)
	D trials	Review carefully, recommended to reject
	F trials	Reject

Definition of abbreviations: EELV = end expiratory lung volume; LCI = lung clearance index; FRC = functional residual capacity; V_T = tidal volume; CO_2 = end tidal concentration of carbon dioxide; N_2 = end tidal concentration of nitrogen. Target range for V_T defined as 10-15ml/kg (represents green line in flow-volume figure). For our quality control criteria we define the pre-phase as the three breaths preceding the start of the washout. The start of the washout is defined as the first breaths of the washout, and the rest of the washout includes all the remaining breaths of the washout until the end of test criteria is reached (including first three breaths below 2.5% normalized end tidal nitrogen concentration).

Table 2: Patient demographics

	Bern		Zurich	
	Before real-time QC	**After real-time QC**	**Before real-time QC**	**After real-time QC**
Patients (n)	31	32	34	30
Female: n (%)	16 (52)	17 (53)	13 (38)	9 (30)
Age (years)	12.9 (5.2; 18.1)	13.4 (5.6; 18.1)	10 (4.0; 18)	11 (4.2; 18.2)
Weight (kg)	42.4 (19.4; 68)	44.2 (19.9; 74.2)	32.2 (15; 68)	40 (16; 96)
Weight z-score	-0.3 (-1.8; 1.6)	-0.4 (-3.3; 1.7)	-0.2 (-2.2; 2.0)	0.09 (-1.8; 2.8)
Height (cm)	150.2 (110.8; 180)	150.7 (117.4; 183.5)	133.5 (99; 172)	141.8 (99; 177)
Height z-score	-0.1 (-3.4; 1.4)	-0.6 (-4.2; 3.3)	-0.4 (-2.4; 1.7)	-0.07 (-1.3; 1.4)
BMI (m/kg2)	18.2 (14; 25.3)	18.9 (14.4 ; 24.8)	17.2 (13.7; 26)	18.2 (13.7; 28)
BMI z-score	-0.2 (-1.6; 1.7)	-0.01 (-1.61; 1.41)	0.03 (-2.1; 2.3)	0.13 (-2.7; 2.8)
LCI 2.5%	11.0 (7.3; 19)	10.8 (7; 17.3)	11.3 (6.5; 19.4)	9.7 (6; 15.2)
FRC (L)	1.9 (0.9; 3.8)	1.9 (0.8; 3.7)	1.6 (0.7; 3.2)	2.0 (0.6; 3.8)

Data are presented as mean (range) or n (%). Definition of abbreviations: BMI = Body mass index, LCI: Lung clearance index, FRC: Function residual capacity, QC: Quality control. Subjects were all children with CF attending routine care at the CF outpatient clinics in Bern and Zurich and randomly selected for this study. Results presented for LCI and FRC refer to the test occasions accepted by the reviewer.

Table 3: MBW trial and test acceptability

	Bern		Zurich	
	Before real-time QC	**After real-time QC**	**Before real-time QC**	**After real-time QC**
Test occasions (n)	31	32	34	30
Test acceptability (at least two acceptable trials)	**18 (58)**	**24 (75)**	**13 (38)**	**21 (70)**
Tests with 2 acceptable trials	11(61)	11(46)	2 (15)	9 (43)
Tests with 3 or more acceptable trials	7 (39)	13 (54)	11(85)	12 (57)

Abbreviations: QC: quality control. Total Number of test occasions evaluated. For a valid test occasion, a minimum of 2 acceptable trials is required. Percentages of accepted tests refer to the total number of tests collected. Percentages of tests with 2 or more trials are referring to the number of accepted tests.

Table 4: Reasons for trial exclusion before implementing real-time quality control

	Bern	Zurich
Before real-time QC		
Trials evaluated	89	97
Trials rejected	38 (42)	53 (55)
F grade trials	16 (42)	36 (68)
End of test criteria not met	7 (44)	4 (11)
Leak during washout	6 (37)	29 (81)
Start of test criteria not met	3 (19)	3 (8)
D grade trials	22 (58)	17 (32)

Abbreviations: QC: quality control. Number of trials evaluated before real-time quality control was implemented with reasons provided for trial exclusion. Percentage of trials with D and F grade refers to the total number of trials rejected. F grade trials were excluded as not meeting technical acceptability criteria, percentage given refers to the total number of F grade trials. D grade trials were of questionable quality and therefore excluded for test interpretation.

Table 5: Reasons for trial exclusion after implementing real-time quality control

	Operator	Reviewer
Bern		
Trials evaluated	96	96
Trials rejected	27 (28)	30 (31)
F grade trials	18 (67)	18 (60)
End of test criteria not met	10 (56)	10 (56)
Leak during washout	8 (44)	8 (44)
Start of test criteria not met	0 (0)	0 (0)
D grade trials	9 (33)	12 (40)
Zurich		
Trials evaluated	91	91
Trials rejected	32 (35)	30 (33)
F grade trials	18 (56)	19 (63)

End of test criteria not met	4 (22)	5 (26)
Leak during washout	14 (78)	14 (74)
Start of test criteria not met	0 (0)	0 (0)
D grade trials	14 (44)	11 (37)

Abbreviations: QC: quality control. Number of trials evaluated after real-time quality control was implemented with reasons for trial exclusion provided for the operator and the reviewer in Bern and Zurich, respectively. Percentage of trials with D and F grade refers to the total number of trials rejected. F grade trials were excluded as not meeting technical acceptability criteria, percentage given refers to the total number of F grade trials. D grade trials were of questionable quality and therefore excluded for test interpretation.

Figures

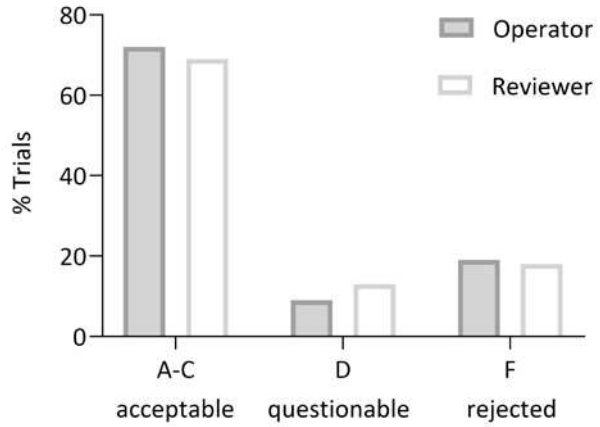

Figure 1a: Agreement in trial grading between operator and reviewer after implementing quality control guidelines in Bern. A total number of 96 trials were evaluated in Bern; Agreement for trial grading into acceptable, questionable, and rejected was 88% (kappa 0.7, $p < 0.001$). A-C grade trials reflect acceptable trials, D grade trials are questionable quality, and F Grade are technically not acceptable.

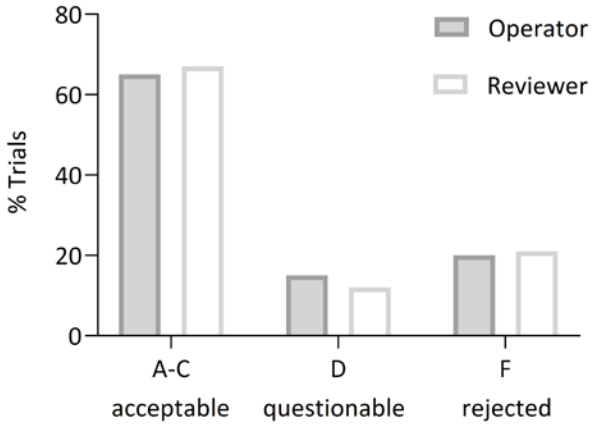

Figure 2b: Agreement in trial grading between operator and reviewer after implementing quality control guidelines in Zurich. A total number of 91 trials were evaluated in Zurich; Agreement for trial grading into acceptable, questionable, and rejected was 86% (kappa 0.7, $p<0.001$). A-C grade trials reflect acceptable trials, D grade trials are questionable quality, and F Grade are technically not acceptable.

Online supplemental material

Real-time quality control guidelines

We propose to assess N_2MBW measurements systematically at three different phases of the measurement. Initially, we examine the 'pre-phase', which includes all the breaths prior to the washout and provides the baseline for the measurement. We define the pre-phase as the three breaths immediately preceding the start of the washout. Second, we examine the 'start of the washout', which we define as the first three breaths of the washout. Finally, we examine the 'rest of the washout', which includes all the remaining breaths of the washout until the end of test criteria is reached (including the first three breaths below 2.5% normalized end tidal nitrogen concentration).

The pre-phase and start of the washout are important for the calculation of outcomes FRC and LCI. The FRC is calculated as the net volume of exhaled N_2 divided by the difference in end tidal N_2 concentration from the start ($CetN2_{start}$) to the end of the washout ($CetN2_{end}$). The LCI is calculated as the cumulative expired volume (CEV) divided by FRC (Equation 1).

$$FRC = \frac{net\ volume\ of\ exhaled\ N2}{CetN2_{start} - CetN2_{end}} \qquad LCI = \frac{CEV}{FRC}$$

The pre-phase is used to calculate the $CetN_2$start and is therefore important for the calculation of FRC. The net volume of exhaled N_2 is calculated based on the integration of N_2 concentration with respect to flow. As the first three breaths of the washout have the largest concentrations of exhaled N_2 they have the biggest influence on the net volume of exhaled N_2. Therefore, the start of the washout is important for the calculation of

both FRC and LCI. Alterations from regular tidal breathing during the pre-phase and start of washout can have a large influence on outcomes and are dealt with more strictly in our criteria.

Supplemental Table E1: Definition of how our quality control criteria differ from the ATS/ERS consensus statement criteria

	ATS/ERS Consensus statement	Additional (+) / modified (*) criteria in our matrix
Evidence of leak	A sudden spike in N_2-concentration during inspiration; premature rise in N_2 signal in early expirogram where N_2 should be zero; Decrease in airway dead space volume; A sudden step change in volume trace or step-up in N_2- concentration plotted vs. lung turnover (TO)	+ Loss of decay or prolonged plateau in end-tidal concentration of N_2; + Concentration of N_2 does not return to zero on inspiration
End of test criteria	Three consecutive breaths where the normalized end-tidal concentration of N_2 fell below 2.5%	+ All three breaths should be tidal breaths
Start of test criteria	Sufficient interval between runs when using resident inert gases to allow inert gas concentration to return to baseline values	+ End-tidal concentration of N_2 for last three breaths of pre-phase ≥77%
Pre-phase (last three breaths of pre-phase)		
Tidal Volume (V_T)	Stable V_T and EELV over the preceding 30 s prior to start washout; No small volume breath immediately prior to start of washout	* Tidal volume stable and relaxed; no sigh, cough, breath-hold, or small breath; small swallow/hesitation acceptable
End expiratory lung volume (EELV)	Deviation in EELV at start of washout within 10% of mean V_T of 5 breaths immediately preceding the start of washout	* No step change in EELV; stable minimal to moderate drift is acceptable
Start of washout (first three breaths of washout):		
Tidal Volume (V_T)	Tidal volume stable and relaxed; regular breathing pattern	* Tidal volume stable and relaxed; no sigh, cough, breath-hold, or small breath; small swallow/hesitation acceptable
End expiratory lung volume (EELV)	Stable volume drift is acceptable; A sudden step change in volume time trace is acceptable provided leak was ruled out	* No step change in EELV; stable minimal to moderate drift is acceptable
Rest of washout:		
Tidal Volume (V_T)	Tidal volume stable and relaxed; regular breathing pattern	* Tidal volume can be moderately variable; Abnormal breaths (sigh, cough, breath-hold) with no leak or release of trapped gas are acceptable
End expiratory lung volume (EELV)	Stable volume drift is acceptable; EELV is stable during washout; A sudden step change in volume time trace is acceptable provided no leak	* Step change with no shift in N_2 signal acceptable
Flow	No coughing	* No evidence of forced exhalation, cough with flow exceeding 1000 ml/s, highly erratic flow

Trapped Gas Release Respiratory rate	No evidence of significant trapped gas release with larger breaths.	+ No evidence of significant trapped gas release (increased end tidal N_2 concentration) with sigh or large breath + No evidence of hyper or hypoventilation
Repeatability	FRC within 25% of the median FRC of all technically acceptable trials	

Definition of our quality control criteria (for acceptable trials) and how it differs from the ATS/ERS consensus statement criteria The symbol '+' indicates additive criteria further to the ATS/ERS consensus statement criteria. The symbol '*' indicates modified criteria to the ATS/ERS consensus statement criteria. Definition of abbreviations: EELV = end expiratory lung volume; LCI = lung clearance index; FRC = functional residual capacity; V_T = tidal volume; CO_2 = end tidal concentration of carbon dioxide; N_2 = end tidal concentration of nitrogen; TO = lung turnover. Target range for V_T defined as 10-15ml/kg (represents green line in flow-volume figure). For our quality control criteria we define the pre-phase as the three breaths preceding the start of the washout. The start of the washout is defined as the first breaths of the washout, and the rest of the washout includes all the remaining breaths of the washout until the end of test criteria is reached (including first three breaths below 2.5% normalized end tidal nitrogen concentration).

Supplemental Table E2: Outcomes before and after implementing real-time quality control (Bern)

Before real-time QC	***Clinical Report***	***Reviewer***	***p-value***
LCI 2.5%	11.6 (3.7)	11.0 (3.4)	0.5
FRC (L)	1.9 (0.7)	1.9 (0.8)	1.0
LCI 2.5% CV	5.2 (3.2)	5.5 (3.2)	0.7
FRC (L) CV	5.3 (3.7)	4.5 (3.6)	0.5
After real-time QC	***Operator***	***Reviewer***	***p-value***
LCI 2.5%	11.3 (2.9)	10.8 (2.6)	*0.6*
FRC (L)	1.8 (0.7)	1.9 (0.7)	0.8
LCI 2.5% CV	4.5 (3.1)	4.8 (3.3)	0.7
FRC (L) CV	4.2 (3.3)	4.5 (3.9)	0.7

Data are presented as mean (standard deviation) or n (%). Definition of abbreviations: QC: quality control, LCI: lung clearance index, FRC: Functional residual capacity, CV: coefficient of variation (%). The 'before real-time QC' comparison examined differences in the outcomes printed on the clinical report vs the outcomes from the trials deemed acceptable by the reviewer. The 'after real-time QC' comparison examined differences in the outcomes reported prospectively by the operator vs retrospectively by the reviewer.

Supplemental Table E3: Outcomes before and after implementing real-time quality control (Zurich)

Before real-time QC	*Clinical Report*	*Reviewer*	*p-value*
LCI 2.5%	11.3 (3.2)	11.3 (2.9)	0.9
FRC (L)	1.5 (0.7)	1.6 (0.6)	0.6
LCI 2.5% CV	6.9 (4.3)	5.7 (3.2)	0.4
FRC (L) CV	7.0 (6.8)	5.1 (3.4)	0.3
After real-time QC	***Operator***	***Reviewer***	***p-value***
LCI 2.5%	9.8 (2.7)	9.7 (2.6)	0.8
FRC (L)	2.0 (0.8)	2.0 (0.8)	1.0
LCI 2.5% CV	5.2 (3.1)	5.3 (2.9)	0.9
FRC (L) CV	4.8 (2.2)	4.6 (2.0)	0.8

Data are presented as mean (standard deviation) or n (%). Definition of abbreviations: QC: quality control, LCI: lung clearance index, FRC: Functional residual capacity, CV: coefficient of variation (%). The 'before real-time QC' comparison examined differences in the outcomes printed on the clinical report vs the outcomes from the trials deemed acceptable by the reviewer. The 'after real-time QC' comparison examined differences in the outcomes reported prospectively by the operator vs retrospectively by the reviewer.

Supplemental Table E4: Impact on outcomes when including D grade trials

	Excluding D grade trials	*Including D grade trials*	*p-value*
LCI 2.5%	10.3 (0.4)	10.2 (0.4)	0.9
FRC (L)	1.9 (0.11)	1.8 (0.1)	0.5
LCI 2.5% CV	5.0 (0.46)	5.0 (0.41)	0.9
FRC (L) CV	4.5 (0.46)	4.8 (0.45)	0.7

Data are presented as mean (standard deviation). Definition of abbreviations: LCI = lung clearance index, FRC = Functional residual capacity, CV: coefficient of variation (%). Results are pooled for both centers. Results presented compare the mean outcomes from all acceptable test occasions including D grade trials (n = 54) to acceptable test occasions whereby D grade trials were excluded (n = 45).

Supplemental Figure E1: Impact of performing quality control on individual test occasions for LCI

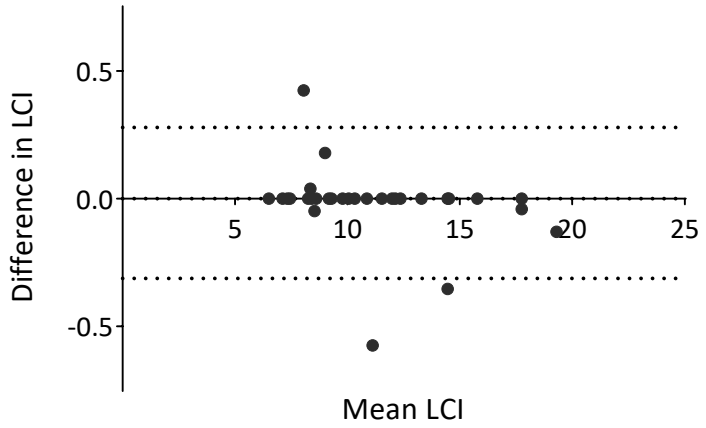

***Figure E1: Bland Altman Plot for the difference in LCI before and after performing quality control*.** We compared the outcomes reported to the clinicians at the time of the measurement before real-time quality control was introduced to the outcomes when test occasions were assessed by the reviewer. While we found no differences in LCI before and after quality control on a population level, the impact of performing quality control for an individual can have a minor impact (+/-0.5 LCI points). Dark grey dots represent mean LCI for each individual, the solid line the mean difference, and the dotted lines the 95% limits of agreement.

Supplemental Figure E2: Impact of performing quality control on individual test occasions for FRC

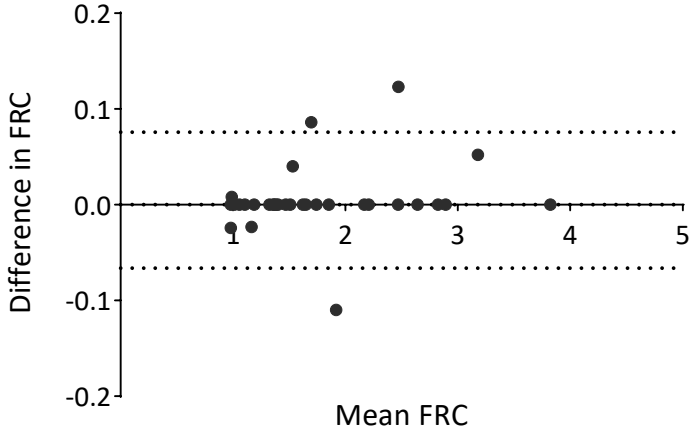

Figure E2*: *Bland Altman Plot for the difference in FRC before and after performing quality control. We compared the outcomes reported to the clinicians at the time of the measurement before real-time quality control was introduced to the outcomes when test occasions were assessed by the reviewer. While we found no differences in FRC before and after quality control on a population level, the impact of performing quality control for an individual can have a minor impact (+/-0.1 Litre FRC). Dark grey dots represent mean FRC for each individual, the solid line the mean difference, and the dotted lines the 95% limits of agreement.